More than 'Mutual Information': Educational and Sectoral Gender Segregation and their Interaction on the Flemish Labour Market

Tom Van Puyenbroeck[♯,♮], Karolien De Bruyne[♯,♮], Luc Sels[♮]
♯: HU Brussel, Stormstraat 2, 1000 Brussel, Belgium; and
♮: KU Leuven, Naamsestraat 69, 3000 Leuven
(corresponding author: tom.vanpuyenbroeck@hubrussel.be)

September 2010

Abstract

We build on the Information Theory foundations of the Mutual Information Index of Segregation [Mora and Ruiz-Castillo, 2003; Frankel and Volij, 2007] to analyze two horizontal dimensions of gender segregation on the labour market. We provide a novel, *three-way* addditive decomposition of their effects on overall segregation. Using survey data from 41,712 Flemish employees, we find that choice of study field has a larger effect on overall segregation than sectoral choice. Their mutual interaction is negative, indicating that sectoral segregation, although low, is still partly explained by educational choices.

1 Introduction

Gender segregation in the labour market and its causes continue to attract the attention of scientists and policy makers. Accurate knowledge about gender segregation builds on micro-data, relating a worker's gender to his or her occupational type, the sector in which that worker is employed, his or her position on the hierarchical ladder, and the like. Information at the individual level is thus key, but is of course too unwieldy as such for, say, monitoring the evolution of gender segregation over time or comparing the extent of gender segregation across different labour markets. This explains both the need for and the frequent use of so-called segregation indices, i.e.

aggregate measures that summarize all relevant individual information and somehow quantify gender segregation in the labour market at large.

The pedigree of segregation indices goes back to the, still popular, dissimilarity index of Duncan and Duncan (1955), but an impressive family of segregation indices has been developed since then.[1] In this paper we focus on the *Mutual Information Index* of segregation. Building on a seminal article by Theil and Finizza (1971) [see also Fuchs, 1975], Mora and Ruiz-Castillo (2003) introduced this index and emphasized its attractive additive decomposability properties. Such a decomposition is necessary if one wants to have a clear and internally consistent idea about the underlying structure of observed segregation (for instance, one may be interested in a regional breakdown, or a blue collar-white collar partitioning, etc.). Frankel and Volij (2007) presented a full ordinal axiomatization for this index, and, by baptizing it the Mutual Information Index, indicated its connection with information theory (going back to Shannon, 1948), from which this index is derived.

Notwithstanding several good theoretical/axiomatic reasons for advocating the Mutual Information index (next to Frankel and Volij, see also Mora and Ruiz-Castillo, 2008), we will here pay attention to its very intuitive interpretation, both conceptually and analytically, in the context of gender segregation. Such an approach may be called for, at least for those who question whether linking the formal concepts of Claude Shannon's information theory to gender segregation is not too farfetched. By focusing on its information theory underpinnings we complement the aforementioned analyses and, in particular, can easily generalize the gender segregation measurement problem to the two-dimensional case.

To understand what we mean by the latter notion, note that accurate knowledge as referred to above indeed renders the gendered division of labour into a multi-dimensional phenomenon, in the sense that several particular divisions of the labour force are conceivable and relevant when measuring gender segregation. Accordingly, gender segregation has many specific appearances, which raises the issue whether and to what extent these are interrelated. Identifying such interrelations is not only important *per se* —i.e., it is instructive in itself to know whether occupational and sectoral segregation, say, are statistically independent or not—, but may also provide an important

[1] The associated theoretical literature is concerned with the most appropriate interpretation of a specific index in its capacity as a summary measure, investigates to what extent specific indices inhibit desirable analytic properties, or, conversely, states a list of specific desirable properties for a segregation measure (i.e., 'axioms') from which specific indices can be derived (see Flückiger and Silber, 1999, for an authoritative introduction to the segregation measurement literature).

input when it comes to designing policies. This is certainly the case for the setting on which we focus here, in which we simultaneously discern sectoral or occupational gender segregation on the one hand, and so-called educational gender segregation on the other.[2] Obviously, different gender policies are appropriate when (a) occupational segregation is largely the combined result of having a strong education/occupation nexus and strong gender biases in the choice of study field, (b) occupational segregation occurs despite a fairly even gender distribution over different educational fiels, or (c) substantial gender biases in education are mitigated once one enters the labour market.

Educational choices traditionally figure among the major factors driving subsequent gender segregation on the labour market, although statistical evidence for their impact today is mixed, at least for EU member states (see e.g. European Commision, 2009; Smyth and Steinmetz, 2008). Arguably however, mixed evidence is precisely what one ought to expect. The interrelation between educational segregation and subsequent labour market segregation may well differ not only over time and between different countries, but also depending on the exact specification of educational categories as well as of relevant labour market divisions. For instance, a quite different message may emerge when one examines the relation between one's 'vertical' choice of educational level (primary, secondary, tertiary schooling) and one's eventual occupation, or between one's 'horizontal'choice of study field (nursing, engineering, arts,...) and one's occupation. Of course, as we focus on a formal measure and its breakdown, it is generic and may be used to study interrelations between two relevant dimensions of gender segregation of many specific kinds. In our empirical section we will illustrate that different messages do emerge, depending on the specific type of data used.

Other papers have addressed the simultaneous measurement of educational and sectoral/occupational gender segregation on the labour market. Valentova, Kristova and Katrnak (2007) measure occupational and educational gender segregation (by field as well as level of study) in 18 European countries. Specifically, they compute the Charles-Grusky (1995) segregation index for each of these three gendered divisions separately,[3] and study the link between occupational segregation and either of the two types of educational

[2]We follow the larger part of the existing literature and focus on *occupational* versus educational segregation in our theoretical part. As far as our empirical application is concerned, we however focus on *sectoral* versus educational segregation. This makes sense as there is a strong statistical correlation between sectoral and occupational segregation (cfr. Section 4).

[3]In our notation, and given $i = 1, ..., I$ educational or occuppational categories, this index is defined as $R = 1/I \sum_i^I R_i$ with $R_i = \left[\ln(p_{female,i}/p_{female}) - 1/I \sum_i^I \ln(p_{female,i}/p_{female})\right]$.

segregation by juxtaposing the respective segregation indices and comparing them across countries. In contrast, this paper employs an *analytical* connection between occupational (or sectoral) and educational segregation indices, which is established via the joint distribution of both characteristics and an associated 'total segregation measure' (with educational and occupational indices being based on the respective marginal distributions of gender types over these two categories). A similar construct appears in Borghans and Groot (1999), who study the trajectory from educational 'presorting' to observed occupational or sectoral segregation and apply this framework to Dutch data (see also Sookram and Strobl, 2009). However, they use the Karmel-MacLachlan index (1988) as a point of departure, which inhibits some less desirable properties, and entails a somewhat intricate decomposition of the trajectory from eductional segregation, over total segregation (via counting 'additional segregation' and correcting for 'reintegration'), to occupational/sectoral segregation (see also Flückiger and Silber (1999, p. 135-137), who additionally discuss a multidimensional extension of the Gini segregation measure).[4,5] Finally, and as we will explain in more detail below, Mora and Ruiz-Castillo (2003) did address the two-dimensional segregation case when using the Mutual Information index , but provide a different ('two-stage') decomposition of total segregation than ours.

The 'three-way' alternative that we propose in section 3, and that we apply to a large sample survey of the Flemish labour market in section 4, preserves the additive structure of the underlying index in that it considers total segregation as the sum of (i) occupational gender segregation, (ii) educational gender segregation, and (iii) the *interaction* of both segregation sources. Positive interaction points to synergetic effects between the two components. Negative interaction indicates (partial) redundancy, or mutually tempering effects, of both components to explain overall segregation. In

[4]It is for instance well-known that the value of this index will not change if there is a 'transfer' of female workers between two occupations when both have either a lower or a higher gender ratio than the overall gender ratio. See e.g. Flückiger and Silber (1999), or Hutchens (2001). As pointed out by the former authors, this less desirable property is also incorporated by the Duncan and Duncan index, of which the Karmel-Maclachlan index is a simple transformation.

[5]In our notation, and with θ a scaling parameter, they measure educational segregation as $ES = \theta \sum_i^I \left| \frac{p_{female,i}}{p_{female}} - \frac{p_{male,i}}{p_{male}} \right|$; occupational segregation as $OS = \theta \sum_j^J \left| \frac{p_{female,j}}{p_{female}} - \frac{p_{male,j}}{p_{male}} \right|$, and total segregation as $TS = \theta \sum_i^I \sum_j^J \left| \frac{p_{female,i,j}}{p_{female}} - \frac{p_{male,i,j}}{p_{male}} \right|$. In going from ES to TS and from TS to OS, Borghans and Groot introduce measures for additional segregation and reintegration, which are to some extent intuitive, but are ultimately explicable by the need to eliminate the absolute operators in explaining the observed transition from ES to OS.

fact, and as we clarify in the next two sections, this three-way decomposition naturally emerges once the information theory underpinnings of the Mutual Information Index explicitly re-enter the analysis.

2 Deconstructing the Mutual Information Index of Segregation

In information theory, the mutual information, usually denoted as $I(X;Y)$ measures how much information about one variable X is revealed on average by knowledge of the other variable Y. It is easy to appreciate why this concept found its way to segregation measurement: if knowledge about a worker's occupation reveals a lot on average about that worker's gender (or *vice versa*), then surely some segregation takes place. In this section we take a closer look at this measure, and discuss some of its equivalent formulations in the segregation context.

2.1 Segregation along one dimension

We start with a simple setting in which segregation between groups is analyzed along only one relevant categorical dimension. Classical examples are the segregation of students belonging to different ethnical groups over different schools in a district or the occupational segregation of male and female workers. Thus, introducing notation, we are concerned with one segregation-relevant dimension $\mathbf{O}$, comprising the categories O_j $(j = 1, .., J)$. There are T individuals in total, where each individual belongs to exactly one of these categories and, moreover, also belongs to a particular group $g \in \mathbf{G}$. In the gender segregation case one evidently has $\mathbf{G} = \{female, male\}$.[6]

Occupational gender segregation implies that gender groups are distributed unevenly over the J occupational categories. When interpreted as a lack of 'evenness', this implies in turn that knowledge about the distribution of *individuals* over different categories is not the same as knowledge about the *gender* distribution over these categories. Operationalizing this definition requires knowledge about the relative frequency distributions of (gender) groups over the different (occupational) categories. Thus, if T_j^g is the number of individuals belonging to group g that have occupation j, we denote by

[6] Although we could have used a notation referring to groups with size F and M such as e.g. in Mora and Ruiz-Castillo (2003), the generic notation $g \in G$ makes clear from the outset that the Mutual Information index of segregation is, by construction, perfectly suited for measuring segregation in a multi-group setting.

$p_{g,j} = P(G = g, O = O_j) = \frac{T_j^g}{T}$: the individuals of group g with occupation j, as a proportion of the total number of individuals,

$p_g = P(G = g) = \frac{\sum_{j=1}^{J} T_j^g}{T} = \sum_{j=1}^{J} p_{g,j}$: the proportion of individuals belonging to group g,

$p_j = P(O = O_j) = \frac{\sum_{g \in G} T_j^g}{T} = \sum_{g \in G} p_{g,j}$: the proportion of individuals belonging to occupation j,

$p_{g|j} = P(G = g | O = O_j) = \frac{T_j^g}{\sum_{g \in G} T_j^g} = \frac{p_{g,j}}{p_j}$: the proportion of individuals in occupation j that belong to group g, and

$p_{j|g} = P(O = O_j | G = g) \frac{T_j^g}{\sum_{j=1}^{J} T_j^g} = \frac{p_{g,j}}{p_g}$: the proportion of individuals belonging to group g that have occupation j.

A Mutual Information Index $I(G;O)$ can be derived from these primitive data in multiple ways, all having a strong intuitive appeal. A first approach builds on gauging the extent to which the actual distributions of both O and G in the labour force diverge from a reference distribution associated with the absence of occupational gender segregation. A common way of measuring the difference between two probability distributions P and Q is via the Kullback-Leibler divergence, defined for discrete distributions as $D_{KL}(P\|Q) \equiv \sum_i P(i) \log \frac{P(i)}{Q(i)}$. Based on this well-established measure, Mora and Ruiz-Castillo (2009) discern three equivalent formulations of the Mutual Information index $I(G;O)$:

$$I(G;O) = \sum_{j=1}^{J} \sum_{g \in G} p_{g,j} \log \left(\frac{p_{g,j}}{p_g p_j} \right) = D_{KL} \left(P\left(G,O\right) \| P(G)P(O) \right), \quad (1)$$

$$= \sum_{g \in G} p_g \sum_{j=1}^{J} p_{j|g} \log \left(\frac{p_{j|g}}{p_j} \right) = \sum_{j=1}^{J} p_g \cdot D_{KL} \left(P\left(O|\, G\right) \| P(O) \right) \quad (2)$$

$$= \sum_{j=1}^{J} p_j \sum_{g \in G} p_{g|j} \log \left(\frac{p_{g|j}}{p_g} \right) = \sum_{g \in G} p_j \cdot D_{KL} \left(P\left(G|\, O\right) \| P(G) \right) \quad (3)$$

The first expression regards occupational gender segregation as the extent to which occupations and gender are statistically dependent. Hence, it gauges

the difference between the actual joint distribution of both characteristics in the labour force (the $p_{g,j}$'s), and the reference distribution $p_g \cdot p_j$ indicating their statistical independency. Expression (2) captures the 'unevenness'-interpretation of occupational segregation; conditioning on gender makes a difference, on average, when assessing the distribution of individuals over occupations. Occupational segregation therefore shows up as an expected positive divergence between the conditional occupational distribution (the $p_{j|g}$'s) and its unconditional reference counterpart captured by the p_j's. Since (1) implies that $I(G;O)$ is symmetric in G and O, one can reverse the roles of occupations and groups to arrive at (3). This formalizes a dual notion of 'unevenness' (*across* occupations), to wit, insufficient 'representativeness' (*of* occupations) (see Frankel and Volij, 2007; Massey and Denton, 1988). Specifically, occupational gender segregation implies the existence of occupational categories with female/male worker ratios that are distinct from the overall gender composition of the labour force. In this case the gender composition of the labour force at large, i.e. the reference distribution in this dual interpretation, diverges from the gender composition of different occupations. Expression (3) is geared exactly towards this interpretation. Each alternative provides a natural interpretation for $I(G;O) = 0$, i.e., complete coincidence of the actual distribution and the no-segregation reference distribution.

Following Mora and Ruiz-Castillo (2009) we thusfar regarded (1), (2) and (3) as well-established statistical divergence measures. Yet the Kullback-Leibler divergence $D_{KL}(P\|Q)$ —alternatively labeled 'relative entropy', or 'information gain'— also appears in information theory, notably as an indicator of the increase in information obtained by using the ('true') distribution P rather than the ('theoretical', 'approximative') distribution Q. Interpreted as such, the three alternative specifications of $I(G;O)$ also bear a very logical connotation. Still, in information theory the mutual information concept is often introduced differently. In the context at hand, it can be taken as the average reduction in uncertainty about a variable such as a worker's gender when one knows that worker's occupation. Formally, one starts from the Shannon entropy information measure

$$H(G) = \sum_{g \in G} p_g \log\left(\frac{1}{p_g}\right),$$

which measures the average uncertainty inherent in the distribution of males and females over the labour force. In a similar fashion one defines the gender

entropy within a specific occupation O_j:

$$H(G|\, O = O_j) = \sum_{g \in G} p_{\,g|j} \log \left(\frac{1}{p_{\,g|j}} \right).$$

If the gender distribution in O_j exactly mirrors the overall gender distribution in the labour force, then O_j is uninformative in the sense that $H(G) = H(G|\, O = O_j)$. Extending to all occupations and providing an expected value, we get the *conditional entropy*:

$$H(G|\, O) = \sum_{j=1}^{J} p_j H(G|\, O = O_j) = \sum_{j=1}^{J} p_j \sum_{g \in G} p_{\,g|j} \log \left(\frac{1}{p_{\,g|j}} \right).$$

Since mutual information is defined as the reduction in uncertainty about G if we know the realization of O, we obtain:

$$I(G;O) = H(G) - H(G|\, O), \tag{4}$$

which is exactly how Frankel and Volij (2007) defined the Mutual Information Index in the context of ethnical segregation[7].

Finally, using $p_g = \sum_{j=1}^{J} p_{gj}$:

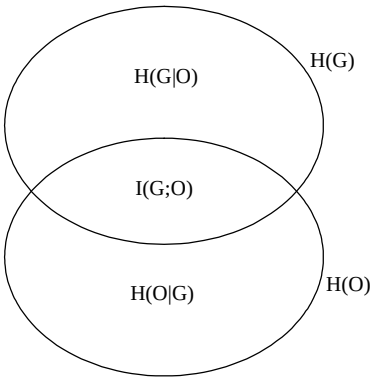
$$\begin{aligned} H(G) - H(G|\, O) &= \sum_{g \in G} p_g \log \left(\frac{1}{p_g} \right) - \sum_{j=1}^{J} p_j \sum_{g \in G} p_{\,g|j} \log \left(\frac{1}{p_{\,g|j}} \right) \\ &= \sum_{j=1}^{J} p_j \sum_{g \in G} p_{\,g|j} \log \left(\frac{p_{\,g|j}}{p_g} \right) = (3), \end{aligned}$$

so indicating the equivalence between (1), (2), (3), and (4).

It is instructive to summarize the concepts hitherto surveyed in a so-called information diagram, which provides set-theoretic representations of information measures (see e.g. Yeung, 2008). The sets $H(G)$ and $H(O)$ refer to the entropy of gender and occupations in the labour force taken separately. The average uncertainty about someone's gender that is unexplained given knowledge of his or her occupation is the conditional entropy $H(G|\, O) \subseteq H(G)$. Consequently, the part of the gender distribution that is explained by knowing the distibution of occupations is $I(G;O) = H(G) - H(G|\, O)$.

[7] In their words (p.15): "The mutual information index equals the entropy of the district's ethnic distribution [the labor market's gender distribution] minus the average entropy of the ethnic [gender] distributions of its schools [over all J occupations]." See also equation (5) in Mora and Ruiz-Castillo (2003) and the way it is subsequently interpreted.

Figure 1: Information diagram for occupational segregation



The information diagram illustrates that $I(G;O)$ is symmetric (and thus, in particular, $I(G;O) = H(O) - H(O|\,G)$; occupational gender segregation is the part of the occupational distribution of workers that is explained by their gender). For our purposes, the diagram's main advantage is that it can easily be extended to the case in which a second segregation dimension enters the analysis.

2.2 Segregation along two dimensions

We now additionally consider the educational group E_i, $i = 1, .., I$ to which a worker from gender group g and with occupation O_j belongs. This case has been studied by Mora and Ruiz-Castillo (2003, 2009), but we reformulate it using information theory concepts. Thus, let $T^g_{i,j}$ be the number of individuals belonging to group g with occupation j and education i. We now have as primitive data:

$p_{g,i,j} = P(G = g, E = E_i, O = O_j) = \frac{T^g_{i,j}}{T}$: the (joint) probability of finding an individual belonging to group g, with occupation j, and with education i, in the labour force,

$p_g = P(G = g) = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} T^g_{i,j}}{T} = \sum_{i=1}^{I}\sum_{j=1}^{J} p_{g,i,j}$: the (unconditional) probability of finding a group g-individual in the labour force,

$p_{g,i} = P(G = g, E = E_i) = \frac{\sum_{j=1}^{J} T^g_{i,j}}{T} = \sum_{j=1}^{J} p_{g,i,j}$: the joint probability of finding

an individual *in the labour force* that belongs to g and has education i,

$p_{g|i} = P(G = g|\, E = E_i) = \frac{\sum_{j=1}^{J} T_{i,j}^{g}}{\sum_{g\in G}\sum_{j=1}^{J} T_{i,j}^{g}} = \frac{p_{g,i}}{p_i}$: the conditional probability of finding an individual *in education category* i that belongs to g,

$p_{g|i,j} = P(G = g|\, E = E_i, O = O_j) = \frac{T_{i,j}^{g}}{\sum_{g\in G} T_{i,j}^{g}} = \frac{p_{g,i,j}}{p_{i,j}}$: the conditional probability of finding a group g-individual among the workers with education i and occupation j,

$p_{g,i|j} = P(G = g, E = E_i|\, O = O_j) = \frac{T_{i,j}^{g}}{\sum_{g\in G}\sum_{i=1}^{I} T_{i,j}^{g}} = \frac{p_{g,i,j}}{p_j}$: the conditional probability of finding a group g-individual, with education i, among the workers with occupation j.

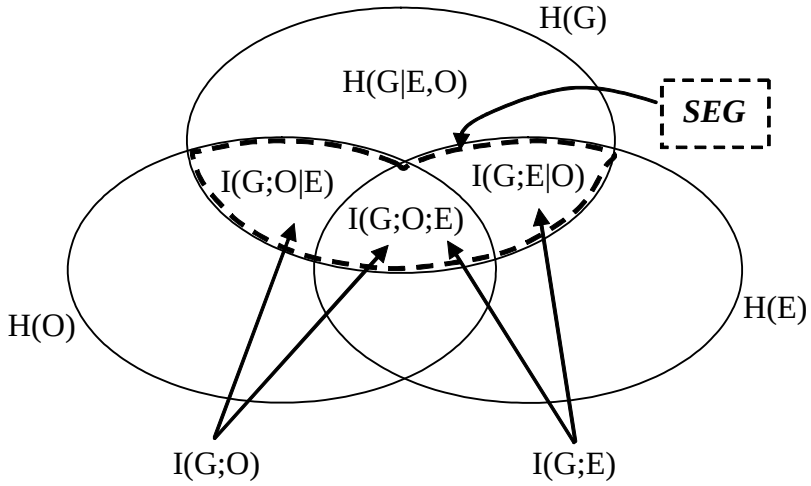
Other unconditional, joint, and conditional probabilities —p_i, p_j, $p_{g,j}$, $p_{i,j}$, $p_{g|j}$, $p_{i|g}$, $p_{g,i|j}$, $p_{i|g,j}$, etc— are similarly defined.

The segregation index in this two-dimensional setting straightforwardly extends the intuition of the mutual information index as given in (4). In our notation, the "overall" measure proposed by Mora and Ruiz-Castillo (2003) is:

$$\begin{aligned} SEG &= \sum_{i=1}^{I}\sum_{j=1}^{J} p_{i,j}\left[\sum_{g\in G} p_{g|i,j}\log\left(\frac{p_{g|i,j}}{p_g}\right)\right] \\ &= \sum_{g\in G}\sum_{i=1}^{I}\sum_{j=1}^{J} p_{g,i,j}\log\left(\frac{p_{g|i,j}}{p_g}\right), \end{aligned} \tag{5}$$

where the expression between square brackets in (5) represents a "*direct* measure of gender segregation" in the group of individuals with education i and occupation j, in relation to the entire labour force. The overall index SEG is a weighted average of these direct measures, with weights based on the relative importance of (i, j)-type workers in the total labour force. We note that (5) extends the information-theoretic structure of the single-

Figure 2: Information Diagram for Two-Dimensional Segregation



dimensional mutual information index $I(G;O)$ as follows:

$$\begin{aligned} SEG &= \sum_{g\in G}\sum_{i=1}^{I}\sum_{j=1}^{J} p_{g,i,j}\log\left(\frac{1}{p_g}\right) - \sum_{g\in G}\sum_{i=1}^{I}\sum_{j=1}^{J} p_{i,j}p_{g|i,j}\log\left(\frac{1}{p_{g|i,j}}\right) \\ &= \sum_{g\in G} p_g \log\left(\frac{1}{p_g}\right) - \sum_{i=1}^{I}\sum_{j=1}^{J} p_{i,j}H(G|\, E=E_i, O=O_j) \\ &= H(G) - H(G|\, E,O). \end{aligned} \tag{6}$$

Hence, in the two-dimensional case segregation boils down to conditioning the uncertainty about the labour force's gender distribution *both* on education and occupation, and subtracting this uncertainty from the one contained in the general unconditional gender distribution $H(G)$. The information diagram in figure 2 (based on Yeung, 2008, p. 61) illustrates the nature of SEG in this extended setting.

Importantly, figure 2 also conveys the message that SEG is *not* a mutual information measure in the strict sense. However, SEG can be readily connected with common Shannon information measures. In particular, the between/within-group decompositions of SEG as proposed by Mora and Ruiz-Castillo (2003) are decompositions of $H(G) - H(G|\, E,O)$ into (a) a proper mutual information index and (b) a so-called *conditional mutual information measure*.

The Mora and Ruiz-Castillo between/within-group decompositions of SEG are either

$$SEG = SEG_{(j)}^{between} + SEG_{(j)}^{within}, \tag{7}$$

or

$$SEG = SEG_{(i)}^{between} + SEG_{(i)}^{within}. \tag{8}$$

In (7) overall segregation is expressed as the combination of (a) an appropriately weighted average of occupational gender segregation, and (b) a weighted average of educational segregation within each of the J occupations. This decomposition is mirrored in (8), combining (a) average gender segregation between educational groups and (b) occupational segregation within each of the I educational groups. In both cases, the between-group component (a) exactly coincides with a mutual information measure as defined above. We show this for $SEG_{(j)}^{between}$, defined by Mora and Ruiz-Castillo as follows (where $T_{\bullet,j}^{\bullet} = \sum_g \sum_i T_{i,j}^g$, $T_{\bullet,\bullet}^g = \sum_i \sum_j T_{i,j}^g$, and $T_{\bullet,j}^g = \sum_i T_{i,j}^g$):

$$SEG_{(j)}^{between} = \sum_{j=1}^{J} \left(\frac{T_{\bullet,j}^{\bullet}}{T} \right) \left[\sum_{g \in G} \frac{T_{\bullet,j}^g}{T_{\bullet,j}^{\bullet}} \log \left(\frac{T_{\bullet,j}^g / T_{\bullet,j}^{\bullet}}{T_{\bullet,\bullet}^g / T} \right) \right].$$

The expression between square brackets captures the divergence of the gender distribution within a specific occupation from the gender distribution in the labour force at large. The factor $T_{\bullet,j}^{\bullet}/T$ is the relative weight of this occupation in the labour force. Evidently, $SEG_{(j)}^{between}$ neutralizes any effect of educational groups. Now, in our notation, and recalling (3):

$$\begin{aligned} SEG_{(j)}^{between} &= \sum_{j=1}^{J} p_j \sum_{g \in G} p_{g|j} \log \left(\frac{p_{g|j}}{p_g} \right) \\ &= I(G;O). \end{aligned} \tag{9}$$

In the same way it can be shown that $SEG_{(i)}^{between} = I(G;E)$.

Combining (6), (7), and (9) with the information diagram in figure 2 subsequently reveals that $SEG_{(i)}^{within} = I(G;E|\,O)$. Similarly, $SEG_{(j)}^{within} = I(G;O|\,E)$. These measures capture conditional mutual information, a Shannon information measure that is related to the bivariate mutual information measure in the same way as conditional entropy is related to the basic entropy measure. We refer to appendix A for a formal proof of the equivalence of $SEG_{(j)}^{within}$ and $I(G;E|\,O)$, resp. $SEG_{(i)}^{within}$ and $I(G;O|\,E)$, but the intuition behind their equivalence is clear. For instance, $SEG_{(j)}^{within}$ measures (aggregate) educational segregation within each of the occupational groups. Thus, within each occupational group, we can measure educational gender

segregation as the extent to which G is conditionally independent from E (compare with (1)):

$$\sum_{g\in G}\sum_{i=1}^{I} p_{g,i|j}\log\left(\frac{p_{g,i|j}}{p_{g|j}p_{i|j}}\right) = I(G;E|\,O=O_j),$$

and we get the conditional mutual information by averaging over the J occupations, i.e.

$$I(G;E|\,O) = \sum_{j=1}^{J} p_j I(G;E|\,O=O_j).$$

Condition mutual information is always non-negative. The minimal value of zero is attained when the conditioning attribute completely explains the association between the other two variables. For example, when $I(G;O|\,E)=0$ the observed occupational segregation is in fact entirely driven by educational choices.[8]

Summing up, using traditional information theory measures we can rewrite SEG as

$$\begin{aligned} SEG &= H(G) - H(G|\,E,O) \\ &= I(G;O) + I(G;E|\,O) \qquad (10) \\ &= I(G;E) + I(G;O|\,E). \qquad (11) \end{aligned}$$

Recalling the intuitive rephrasement at the beginning of this section, (10) means that if knowing a worker's occupation reveals information about that worker's gender, and, furthermore, a worker's education also reveals information about gender when controlling for the occupational effect, then both occupational and educational gender segregation are present in the labour force. In particular, (6), (10) and (11) show how the unidimensional segregation measure $I(G;O)$ is generalized in a two-dimensional setting. In the next section we continue to build on information theory to provide an additional decomposition of SEG.

3 A three-way decompostion of SEG

The commutative roles of E and O in the alternative decompositions , (7)/(10) and (8)/(11) allow isolating either pure occupational segregation $I(G;O)$ or

[8] In the language of Borghans and Groot (1999), in such a case there is only 'educational presorting' on the labour market.

pure educational segregation $I(G;E)$ as a component of overall segregation. While complemented with their proper within-group segregation measure, it is not straightforward to distill a clear-cut overall picture about two-dimensional gender segregation using (7) and (8). Either one reports to what extent G and O are statistically dependent, checking additionaly whether G and E are dependent conditioned on O, or one examines the dependency of G and E, complementing this with a check for the dependency of G and O when conditioned on E. In some cases it may be preferable to integrate both perspectives, so avoiding to decompose overall segregation in this two-stage fashion. Combining (7) and (8) yields

$$\begin{aligned} SEG &= SEG_{(j)}^{between} - (SEG_{(i)}^{between} - SEG_{(j)}^{within}) + SEG_{(i)}^{between} \quad (12) \\ &= SEG_{(j)}^{between} - (SEG_{(j)}^{between} - SEG_{(i)}^{within}) + SEG_{(i)}^{between}. \quad (13) \end{aligned}$$

At first sight (12) and (13) complicate the interpretation of SEG. However, in terms of the concepts displayed in figure 2 these decompositions allow introducing the information measure $I(G;E;O)$:

$$SEG = I(G;O) - I(G;E;O) + I(G;E). \quad (14)$$

The measure $I(G;E;O)$ is known as the *multivariate mutual information.* It is defined as (compare with the middle term in (12) and (13) respectively):

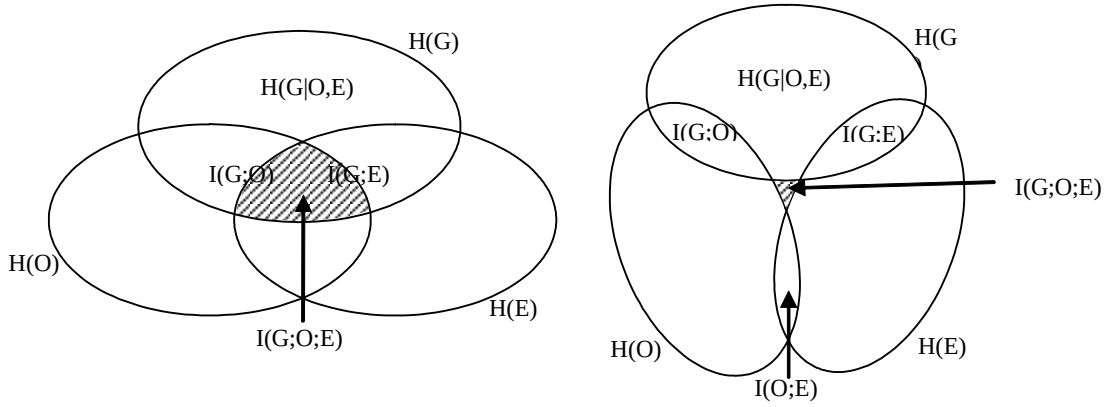
$$\begin{aligned} I(G;E;O) &= I(G;E) - I(G;E|\,O) \\ &= I(G;O) - I(G;O|\,E) \quad (15) \end{aligned}$$

(see e.g. Yeung, 2008, p. 60). Unlike the classic Shannon information measures we have hitherto encountered, $I(G;E;O)$ can be negative. In our setting this may well occur, for example when average occupational segregation as measured within educational groups is larger than occupational segregation as measured when the effect of educational groups is not considered.

In fact, we suggest to focus on $-I(G;E;O)$ rather than $I(G;E;O)$, notably because the former has been interpreted in the literature as the $[G;E;O\text{-}]$ *interaction information* (going back to McGill, 1954; see also Garner and McGill, 1956). Observe that $-I(G;E;O)$ as in (15) reports the extra (average) information about gender carried by knowing the combinations (E_i, O_j) above the information stemming from just knowing the O_j's. Put otherwise, it measures the effect of education on the extent of occupational gender segregation, which evidently is a valuable statistic.[9]

[9] Of course, since $-I(G,E,O)$ is symmetric, it also measures the effect of occupation on educational gender segregation; or the effect of gender on the relation between education and occupation. In general, it is the gain (or loss) in information about the statistical relationship between any two variables due to additional knowledge of the third.

Figure 3: Negative (left) and Positive (right) Interaction



Decomposing SEG as (14) thus involves three terms, two of which are now pure mutual information indices, while the third term captures the interaction between both, i.e.:

total segregation = occupational segregation + educational segregation + their interaction.

Figure 3 (based on Leydesdorff, *forthc.*) illustrates the meaning of the sign of the information interaction term. The left part of Figure 3 depicts the case of negative interaction. In this case total segregation would be overestimated when it is conceived as being the sum of educational and occupational segregation. Hence, there is a (partial) informational overlap, or redundancy, in using both segregation measures when we want to explain gender segregation. Specifically, using for instance (15), the effect of 'pure' occupational gender segregation $I(G;O)$ is (partly) mitigated once the statistical association of education with occupational choices and with gender $I(G;O|\,E)$ is accounted for. To take an extreme example, the gender information contained in the signal that one is dealing with engineers largely coincides with the gender bias in engineering studies and the strong connection between being an engineer and having studied engineering. If such relations characterize the typical education-occupation-gender nexus on a labour market, we are in the left part of figure 3.

Conversely, the right part of Figure 3 indicates synergetic effects of educational and occupational gender segregation; the two partial measures alone do not suffice to explain total segregation. Thus, there is a sufficient number of instances in which knowing someone's education provides extra information (on average) about the strength of association between occupation and gender (i.e., $I(G;O|\,E) > I(G;O)$). For example, it may be difficult to have an idea about someone's gender if we are informed that this person is a teacher, but we may make a better guess if we know that this teacher has a degree in engineering.

Within this perspective, several typologies of a segregated labour market can emerge, both depending on the levels of the constituent mutual information indices and their interaction. Both $I(G;E)$ and $I(G;O)$ may be high (or low), but they may be mutually mitigating or re-inforcing each other. Or, a high level of educational segregation can co-exist with a low level of occupational segregation, indicating that gender biases in education are considerably neutralized once graduates enter the labour market. Negative information interaction in that case additionally conveys that, this neutralization notwithstanding, educational choices still play a role to explain the (low) level of occupational gender segregation that is observed. Against the same background, positive interaction indicates the importance of new gender/education/occupation patterns, over and beyound those associated with educational and occupational segregation separately, as an explanatory factor for the observed difference in the levels of both partial measures. Similar inferences can be made in the opposite case where the extent of occupational segregation is larger than that of educational segregation.

4 Educational and sectoral gender segregation on the Flemish labour market

We illustrate our methodology with an application to the Flemish labour market.[10] We use data from a biannual Salary Survey organised by the Research Centre for Organisation Studies of the Katholieke Universiteit Leuven for 2006. After cleaning the dataset for our purposes, $41,711$ workers remained, of which $17,410$ (41.7%) are female and $24,302$ (58.3%) are male. As there are no administrative data available in Belgium at a sufficiently detailed level, this dataset is the best available for our purpose. As compared

[10] Since in Belgium regional governments are responsible for both educational policy and important aspects of labour market policy, regional data are the most appropriate for an integrated segregation research.

to the labour market at large, and despite the popularity of this survey in Flanders, the selection mechanism leads to an overrepresentation of younger and higher educated people. The average respondent will therefore differ somewhat from the average employee in Flanders. The dataset does not contain blue collar or part-time workers - only white-collar workers and civil servants. Given that some sectors are more prone to offering part-time jobs, and since women engage in part-time work more often than men, this implies that our results are likely to underestimate the true extent of gender segregation on the Flemish labour market at large. For the purpose of illustrating our research question we do not consider this to be a problem. In addition, in the narrower terms of mirroring the educational and sectoral choices of Flemish white-collar workers the sample is actually quite representative, as we now indicate.

Respondents are divided over 46 different educational types and 29 sectoral categories. Tables 3 and 4 in the Appendix show the different categories. Note that, given the nature of our data, we use sectors rather than occupations in this section as one of the segregation-relevant dimensions. We return to this observation below.

The 46 educational types can be largely divided into three groups, reflecting the ternary structure of higher education in Flanders: professional bachelor courses are only provided by colleges of higher education. Some of these colleges and all universities offer academic bachelor and master courses in different fields. In our dataset, 49% of the respondents had a professional bachelor diploma (PB), 17% a college master degree (CM), and 34% a university master degree (U). This sample stratification almost coincides with the actual 2006-2007 student enrollment figures as recorded by the Flemish Ministry, with 78,526 (48.1%) students enrolled in a program leading to a PB degree, 23,951 (14.7%) in the CM category, and 60,866 (37.2%) in the U category.

The 29 sectors can also be divided in different groups: the primary sector (PR), the industry sector (I), the services sector (S) and the public sector (P). Most respondents (46%) work in the services sector, 27% in the industrial sector and 25% in the public sector. The primary sector is negligible – which is not surprising given the very limited importance of the agricultural sector in Flanders. In fact, official Flemish statistics for 2006 reveal that 1% of the working force had a job in the primary sector, 26.1% worked in the Industry sector, and the remaining 78% in the 'services sector' (which includes public sector employees).

We combine all segregation-related information in Tables 1 and 2. The first three columns of Table 1 show the share of women/men/all workers with a specific education (i.e., each row provides a number for $p_{female,i}$, $p_{male,i}$ and

p_i respectively). The fourth column provides per-education female/male ratios, which constitutes our sorting criterion for the rows in Table 1. Thus, women are most overrepresented in tourism (PB11), bio-medical sciences (U21), journalism (CM3), psychology (U8), and nursing studies (PB6). Men on the other hand are overrepresented in military sciences (U22), construction (PB14), computer science (PB3 and U11), civil engineering (U2) and industrial engineering (CM1).

Table 2 provide similar statistics for each sector of employment, i.e. $p_{female,j}$, $p_{male,j}$ and p_j, and per-sector female/male ratios. Women are overrepresented in human resources (S8) and tourism (S9), in the socio-cultural (P7), welfare (P4) and health services (P1) sectors. Men are overrepresented in IT services (S1), the metal industry (I1), the telecommunication (S6), chemical (I2), and construction sectors (I5).

We move on from descriptive statistics to the measurement of gender segregation in columns 5 to 8 of Tables 1 and 2. In Table 1, the bottom line of the fifth column reports pure eductional segregation $I(G;E)$, as computed from its constituent elements in the preceeding rows. The next column reports occupational gender segregation within each educational group, $I(G;O|\,E=E_i)$, leading eventually to $I(G;O|\,E)$ as reported in the last line. Column 7 combines both sources, yielding eventually a total (two-dimensional) gender segregation value of 18.31. Recalling that $I(G;E) + I(G;O|\,E) = I(G;O) + I(G;E|\,O)$, the same value appears at the corresponding position in Table 2, which otherwise contains information relating to $I(G;O)$ and $I(G;E|\,O)$.

Table 1 conveys that a limited number of educational categories contributes a lot to total segregation. The educational category office management-languages (PB1) has a large impact on total segregation for two reasons. It is a large category (20% of all workers) and its female/male composition is definitely irrepresentative for the average female/male ratio. As for the contribution of male overrepresented categories, we note four educational groups with a large impact: industrial engineer (CM1), computer science (PB3), civil engineer (U2) and industrial sciences (PB2).

Concerning the relation between sector and gender, Table 2 shows that the health (P1) and education sector (P2) contribute significantly to total segregation. Both are fairly large sectors, in which women are overrepresented. Other sectors with a high segregation are the metal (I1) and the IT sector (S1). Again these are quite large, and here men are overrepresented. Note that the federal government (P3) and international government and institutions (P8) are almost gender neutral (their between-sectors segregation

Table 1: Descriptive Statistics and Gender Segregation, Flanders, 2006: Educational Partition

	Labour force distribution across educations (%)			% females by education	Gender segregation Indices			
Education	Female	Male	Total		Between	Within	Total	% of total Segregation
PB11	1,63	0,30	0,86	79,55	0,369	0,071	0,440	2,40
U21	0,41	0,09	0,22	77,17	0,083	0,045	0,128	0,70
CM3	2,96	0,72	1,66	74,53	0,529	0,076	0,605	3,31
U8	3,42	1,06	2,05	69,79	0,474	0,075	0,548	3,00
U7	3,70	1,20	2,25	68,84	0,484	0,094	0,578	3,16
PB6	3,85	1,27	2,34	68,51	0,493	0,109	0,602	3,29
PB17	0,24	0,08	0,14	68,33	0,030	0,048	0,078	0,42
PB7	2,45	0,81	1,50	68,27	0,309	0,095	0,404	2,21
U15	0,94	0,35	0,59	66,13	0,103	0,044	0,147	0,80
PB4	5,89	2,20	3,74	65,75	0,630	0,135	0,765	4,18
PB13	0,48	0,20	0,32	63,64	0,044	0,057	0,101	0,55
PB8	2,65	1,16	1,78	62,18	0,217	0,085	0,302	1,65
U10	1,44	0,73	1,02	58,55	0,084	0,084	0,168	0,92
PB1	27,63	14,93	20,23	57,00	1,373	0,542	1,915	10,46
U13	0,82	0,46	0,61	55,69	0,035	0,100	0,135	0,74
CM4	1,46	0,89	1,13	54,03	0,050	0,073	0,123	0,67
U3	4,92	3,01	3,80	53,94	0,165	0,134	0,299	1,63
PB5	3,16	2,01	2,49	52,98	0,092	0,080	0,172	0,94
CM6	0,30	0,22	0,25	50,00	0,005	0,038	0,043	0,23
U20	0,32	0,24	0,27	48,67	0,004	0,027	0,030	0,17
U6	2,70	2,20	2,41	46,81	0,018	0,113	0,131	0,71
U14	0,62	0,51	0,56	46,35	0,003	0,053	0,057	0,31
PB12	0,29	0,27	0,28	43,10	0,000	0,040	0,040	0,22
U18	0,19	0,19	0,19	42,31	0,000	0,040	0,040	0,22
PB10	0,98	0,98	0,98	41,67	0,000	0,171	0,171	0,93
U17	1,01	1,01	1,01	41,67	0,000	0,106	0,106	0,58
U4	2,61	2,63	2,62	41,58	0,000	0,107	0,107	0,58
PB9	0,67	0,67	0,67	41,43	0,000	0,081	0,081	0,44
U12	0,59	0,62	0,60	40,48	0,000	0,048	0,049	0,27
PB16	0,22	0,24	0,23	39,18	0,000	0,042	0,042	0,23
CM2	3,42	3,90	3,70	38,56	0,011	0,160	0,171	0,93
CM5	0,52	0,60	0,57	37,97	0,002	0,132	0,135	0,74
U16	1,37	1,66	1,54	37,17	0,010	0,054	0,064	0,35
U9	1,37	1,69	1,56	36,83	0,011	0,106	0,117	0,64
U19	0,10	0,13	0,12	35,42	0,001	0,014	0,015	0,08
U1	3,64	5,20	4,55	33,39	0,096	0,146	0,242	1,32
U5	1,82	3,24	2,64	28,65	0,140	0,076	0,216	1,18
PB18	0,08	0,16	0,12	26,92	0,009	0,027	0,035	0,19
PB2	2,62	9,32	6,53	16,79	1,353	0,347	1,700	9,29
PB15	0,03	0,12	0,08	14,29	0,022	0,021	0,042	0,23
U11	0,42	1,93	1,30	13,47	0,356	0,041	0,397	2,17
CM1	2,98	14,71	9,81	12,66	2,869	0,306	3,174	17,34
U2	1,26	6,28	4,19	12,60	1,229	0,212	1,441	7,87
PB3	1,79	9,32	6,18	12,07	1,890	0,101	1,991	10,88
PB14	0,06	0,41	0,26	9,17	0,100	0,029	0,129	0,70
U22	0,01	0,09	0,06	4,35	0,030	0,001	0,031	0,17
TOTAL	**100**	**100**	**100**	**41,74**	**13,724**	**4,582**	**18,306**	**100**

Table 2: Descriptive Statistics and Gender Segregation, Flanders, 2006: Sectoral Partition

	Labour force distribution across sectors (%)			% female by sector	Gender segregation indices			
Sector	Female	Male	Total		Between	Within	Total	% of total Segregation
S8	4,07	1,31	2,46	68,94	0,535	0,207	0,742	4,05
S9	1,71	0,65	1,09	65,13	0,175	0,145	0,320	1,75
P7	2,08	0,87	1,37	63,18	0,184	0,084	0,268	1,47
P4	4,73	1,99	3,13	63,05	0,415	0,328	0,743	4,06
P1	8,75	3,97	5,96	61,21	0,659	0,505	1,165	6,36
P2	9,39	4,67	6,64	59,03	0,578	0,580	1,158	6,32
S7	5,03	3,14	3,93	53,42	0,156	0,285	0,441	2,41
Z1	1,68	1,12	1,35	51,95	0,041	0,296	0,337	1,84
P5	3,15	2,20	2,59	50,65	0,060	0,296	0,356	1,94
I7	0,88	0,64	0,74	49,84	0,014	0,082	0,096	0,52
I3	1,42	1,04	1,20	49,60	0,022	0,206	0,227	1,24
S10	3,97	2,97	3,39	48,97	0,052	0,431	0,483	2,64
PR1	0,25	0,20	0,22	47,83	0,002	0,050	0,053	0,29
P6	2,12	1,67	1,86	47,67	0,019	0,204	0,224	1,22
S3	8,47	7,11	7,68	46,03	0,042	1,205	1,247	6,81
P8	0,85	0,72	0,77	45,82	0,004	0,173	0,177	0,97
P3	4,12	4,11	4,12	41,79	0,000	0,491	0,491	2,68
S5	2,85	2,88	2,87	41,52	0,000	0,284	0,284	1,55
S2	6,81	7,74	7,35	38,64	0,021	0,454	0,475	2,60
S11	1,72	2,05	1,92	37,55	0,010	0,323	0,334	1,82
S4	4,15	5,02	4,66	37,18	0,029	0,442	0,471	2,57
I4	2,42	3,01	2,76	36,51	0,023	0,271	0,294	1,60
I6	1,48	1,95	1,75	35,21	0,023	0,299	0,321	1,75
I8	1,00	1,53	1,31	31,81	0,040	0,314	0,354	1,93
I5	2,19	3,50	2,95	30,93	0,106	0,492	0,598	3,27
I2	3,64	6,23	5,15	29,53	0,237	0,700	0,937	5,12
S6	2,27	4,59	3,62	26,21	0,274	0,613	0,887	4,84
I1	4,72	11,20	8,50	23,19	0,932	1,561	2,493	13,62
S1	4,07	11,92	8,64	19,66	1,374	0,959	2,332	12,74
TOTAL	**100**	**100**	**100**	**41,74**	**6,026**	**12,280**	**18,306**	**100%**

value being close to 0). Apparently, both governmental levels follow an equal opportunity policy in the sense of mirroring the overall gender composition of the labour force.

A two-stage within/between group decomposition of total segregation (18.31) *à la* Mora and Ruiz-Castillo is evidently possible on the basis of these data. Pure educational segregation ($I(G;E)$) is 13.72, while its conditional occupational complement ($I(G;O|\,E)$) is 4.58. In other words, knowing the education of a person in our sample will on average be more informative about that person's gender than knowledge of his or her sector of employment (when one knows the educational group of that person). There will, for instance, be more segregation between people with an educational background in engineering and between people with a background in nursing, than between engineers working in the health sector and engineers working in the IT sector.[11] Conversely, our data reveal a low level of pure sectoral segregation ($I(G;O)$ is 6.03) and a higher average level of educational segregation within sectors ($I(G;E|\,O)$ is 12.28); knowing the sector of employment will determine the gender less than knowing one's education, given the information about that person's sector. There is, for instance, less segregation between people in, say, health services and the construction sector than between people in the construction sector that studied engineering or office management. The results for the consulting and R&D sector (S3) can be taken as a good illustration of the average tendencies. Its pure sectoral segregation component is very small, indicating this to be a fairly gender-neutral sector. But within the sector there are large education-related differences. This might be due to the fact that a lot of (male) engineers work in research and development and a lot of (female) office managers in consulting. Again, since this sector is fairly large, it contributes substantially to observed total segregation.

Finally, we look at the results of our three-way decomposition. For our data we obtain

$$\begin{aligned}\text{total segregation} = \quad & \text{sectoral segregation} + \text{educational segregation} \\ & + \text{their interaction},\end{aligned}$$

[11] In the main text we only focus on horizontal (fields of education) and not vertical (level of education) educational segregation. To the extent that a hierarchy in the three educational groups can be perceived (professional bachelors, college masters and university masters), and performing a classic (one-dimensional) group decomposition of educational segregation, we find that educational segregation between these 3 groups (1.33) is much smaller than educational segregation within these groups (12.39). Since this categorization explains educational segregation to a limited extent, we focus on the relation between choice of study field and sectoral segregation.

hence:

$$18.31 = 6.03 + 13.72 + (-1.44).$$

The values for sectoral segregation and educational segregation are qualitatively in line with those of other recent studies, based on different datasets and using different measures. For example, the European Commission's (2009) report on gender segregation, finds that Belgium (i.e. the Flemish, Walloon and Brussels regions taken together) only ranks 21st among 29 European countries in terms of sectoral segregation, as measured with the Karmel-Maclachlan index, while it occupies the 14th position in terms of educational segregation (on the basis of the Duncan and Duncan index). Valentova, Krizova and Katrnak (2009) find that Belgium is somewhat a-typical among EU countries, as it combines relatively low occupational segregation with relatively high (horizontal) educational segregation.

Our methodology additionally enables to explore the connection between both sources in an internally consistent manner via the interaction term. For our data we observe negative interaction between sectoral and educational gender segregation on the Flemish labour market. As pointed out above, this means that educational segregation and sectoral segregation values are to some extent mutually redundant in explaining total, i.e. two-dimensional segregation. Put otherwise, educational choices are not only more important than sectoral choices as a 'first order' explanatory factor for the labour force's gender composition, but in fact also partially explain the first order effect of sectoral choices. As stated in the introduction, such a finding might call for a particular gender policy, in this case evidently oriented towards limiting educational segregation.

In this respect, one may argue that focusing on educational segregation is preferable anyhow, if only because it is plausibly far more difficult to develop policies geared directly towards sectoral gender segregation, and given that our decomposition leaves occupational segregation out of the analysis. However, there is a strong statistical correlation between sectoral segregation and occupational segregation, even if values for occupational segregation measures as a rule are higher than for sectoral segregation measures (European Commission, 2009, p. 33-34). Given this connection and the findings of Valentova et al. (2009), the policy recommendation would remain largely unaffected. Furthermore, precisely in view of the negative interaction result we can state that sectoral segregation, although low, is still partly explained by educational choices.

More generally, we here touch upon a point made in the introduction, viz. that the interrelation between measured educational segregation and

subsequent labour market segregation is likely to be different both across countries and depending on the exact specification of the relevant labour market divisions. In particular, this means that negative interaction is not always to be expected. This can be readily demonstrated with the results for the Spanish labour market in 1977 as reported in Mora and Ruiz-Castillo (2003). Their educational criterion is essentially vertical (going from 'low' to 'college' education), and their second criterion is based on occupational groups. In this setting, they find that $I(G;E) = 1.77$ and $I(G;O|\,E) = 28.27$, which, recalling (10), implies a total segregation value of 30.05. Their commutative two-stage decomposition yields $I(G;O) = 27.00$ and $I(G;E|\,O) = 3.04$. Thus, most of the observed gender segregation in their case originates from occupational choices, whether measured directly or within education subgroups. Using (14), we can add that the interaction between pure educational segregation and pure occupational segregation is positive ($30.05 = 1.77 + 27.00 + 1.28$). This indicates that the observed increase 'from' educational segregation 'to' occupational segregation arises from an enhancing effect of educational choices on occupational segregation, an observation that lends support to gender policies that focus on the transition from schooling to the labour market.

5 Conclusion

Educational choices traditionally figure among the major factors driving occupational/sectoral gender segregation on the labour market. Several papers have studied both types of segregation and the possible link between them. Exploiting the foundations of the Mutual Information Index of Segregation, this link is analytically established via a three-way additive decomposition, which exactly introduces the interaction next to two pure segregation (mutual information) indices. Our application to survey data from 41,712 Flemish employees provides an example of the way in which such information is useful, both descriptively and in pointing towards specific gender policies. In our empirical study, sectoral segregation is less important than educational segregation, and, moreover, is still partly explained by educational choices. Both the theory and a second empirical example show that positive interaction, i.e. synergetic effects, between two dimensions of gender segregation are a real possibility in other settings. This observation indicates an obvious avenue for additional empirical research, using the same methodology with data from other labour markets. Finally, although we also leave this for further research, we point out that an essentially similar way of decomposing total segregation is possible using an n-dimensional classification, using

relevant concepts that have been developed in information theory.

Acknowledgement The authors wish to thank Jef Hendrickx for his valuable comments on an earlier version of this paper.

References

[1] Borghans, L. and Groot, L. (1999), Educational Presorting and Occupational Segregation, *Labour Economics* 6, 375-395.

[2] Duncan, O.D., and Duncan, B. (1955), A Methodological Analysis of Segregation Indices, *American Sociological Review* 20, 210-217.

[3] European Commision, Directorate-General Employment, Social Affairs and Equality of Opportunity (2009), *Gender Segregation in the Labour Market: Root causes, Implications and Policy Responses in the EU*, Luxemburg, Publications Office of the European Union, 111 p.

[4] Flückiger, Y., and Silber, J. (1999), The Measurement of Segregation in the Labor Force, Heidelberg–New-York, Physica-Verlag.

[5] Frankel, D.M., and Volij, O. (2007), Measuring Segregation, mimeo, Iowa State University.

[6] Fuchs, V. (1975), A Note on Sex Segregation in Professional Occupations, *Explorations in Economic Research* 2: 105-111.

[7] Garner, W.R. and McGill, W.J. (1956), The Relation between Information and Variance Analysis, *Psychometrika* 21, 219-228.

[8] Hutchens, R.M. (2001), Numerical measures of segregation: desirable properties and their implications, *Mathematical Social Sciences* 42, 13-29.

[9] Karmel, T., and Maclachlan, M. (1988), Occupational Sex Segregation: Increasing or Decreasing, *Economic Record* 64, 187-195.

[10] Leydesdorff, L. (*forthc.*), Interaction Information: Linear and Nonlinear Interpretations, *International Journal of General Systems.*

[11] Luenberger, D. (2006), *Information Science*, Princeton, Princeton University Press.

[12] Massey, D.S., and Denton, N. (1988), The Dimensions of Racial Segregation, *Social Forces* 67, 281-315.

[13] McGill, W.J. (1954), Multivariate Information Transmission, *Psychometrika* 19, 97-116.

[14] Mora, R., and Ruiz-Castillo, J. (2003), Additively Decomposable Segregation Indexes. The Case of Gender Segregation by Occupations and Human Capital Levels in Spain, *Journal of Economic Inequality* 1: 147-179.

[15] Mora, R., and Ruiz-Castillo, J. (2008), A Defense of an Entropy Based Index of Multigroup Segregation, Working Paper 07-76, Economics Series 45, Universidad Carlos III.

[16] Mora, R., and Ruiz-Castillo, J. (2009), The Statistical Properties of the Mutual Information Index of Multigroup Segregation, Working Paper 09-84, Economics Series 43, Universidad Carlos III.

[17] Shannon, C. E. (1948), A Mathematical Theory of Communication, *Bell System Technical Journal* 27, 379–423 & 623–656.

[18] Smyth, E. and Steinmetz, S. (2008), Field of Study and Gender Segregation in European Labour Markets, *International Journal of Comparative Sociology*, 49, 257-281

[19] Sookram, S., and Strobl, E. (2009), The role of Educational Choice in Occupational Gender Segregation: Evidence from Trinidad and Tobago, *Economics of Education Review* 28, 1-10.

[20] Valentova, M., Krizova, I., and Katrnak T. (2007), Occupational Gender Segregation in the Light of Segregation in Education: A Cross-National Comparison, IRISS Working Paper 2007-04, CEPS-Instead.

[21] Yeung, R.Y. (2008), *Information Theory and Network Coding*, Springer Verlag.

A Within-group segregation as conditional mutual information

To facilitate comparison with the original formulation of $SEG_{(j)}^{between}$ by Mora and Ruiz-Castillo (2003, section 2), we substitute $F_{i,j}$ for $T_{i,j}^{g}$ when $g =$

$female$ and $M_{i,j}$ for $T^g_{i,j}$ when $g = male$. Educational gender segregation within a specific occupation requires comparison of the proportions $F_{i,j}/(F_{i,j} + M_{i,j}) \equiv w_{i,j}$ resp. $M_{i,j}/(F_{i,j} + M_{i,j}) \equiv 1 - w_{i,j}$ with the within-occupation averages $\sum_i F_{i,j}/(\sum_i (F_{i,j} + M_{i,j})) \equiv W_j$ resp. $\sum_i F_{i,j}/(\sum_i (F_{i,j} + M_{i,j})) \equiv 1 - W_j$. For any specific education i we have:

$$I^{ji} = w_{i,j} \log\left(\frac{w_{i,j}}{W_j}\right) + (1 - w_{i,j}) \log\left(\frac{1 - w_{i,j}}{1 - W_j}\right),$$

and average educational segregation within the j-th occupation is therefore measured by

$$I^j = \sum_{i=1}^{I} \left(\frac{F_{i,j} + M_{i,j}}{\sum_i (F_{i,j} + M_{i,j})}\right) \left[w_{i,j} \log\left(\frac{w_{i,j}}{W_j}\right) + (1 - w_{i,j}) \log\left(\frac{1 - w_{i,j}}{1 - W_j}\right)\right].$$

or, since $\frac{F_{i,j}+M_{i,j}}{\sum_i (F_{i,j}+M_{i,j})} = p_{i,j}/p_j$, $w_{i,j} = p_{g=female,i,j}/p_{i,j}$, $1-w_{i,j} = p_{g=male,i,j}/p_{i,j}$, $W_j = p_{g=female,j}/p_j$, $1 - W_j = p_{g=male,j}/p_j$:

$$\begin{aligned} I^j &= \sum_{i=1}^{I} \frac{p_{i,j}}{p_j} \sum_{g \in G} \left[\frac{p_{g,i,j}}{p_{i,j}} \log\left(\frac{p_{g,i,j}/p_{i,,j}}{p_{g,j}/p_j}\right)\right] = \sum_{g \in G} \sum_{i=1}^{I} \left[\frac{p_{g,i,j}}{p_j} \log\left(\frac{p_{g,i,j}/p_{i,j}}{p_{g,j}/p_j}\right)\right] \\ &= \sum_{g \in G} \sum_{i=1}^{I} \left[p_{\,g,i|j} \log\left(\frac{p_{g,i,j}/p_j}{p_{g,j}/p_j \cdot p_{i,j}/p_j}\right)\right] \\ &= I(G; E|\, O = O_j). \end{aligned}$$

The total (average) effect of within-occupation educational segregation is then

$$SEG^{between}_{(j)} = \sum_{j=1}^{J} \frac{\sum_{i=1}^{I} (F_{i,j} + M_{i,j})}{T} I^j,$$

or, since the weighting term is p_j:

$$SEG^{between}_{(j)} = \sum_{j=1}^{J} p_j I(G; E|\, O = O_j) = I(G; E|\, O).$$

The fact that $SEG^{between}_{(i)} = I(G; O|\, E)$ is demonstrated similarly ■.

Table 3: Educational Partitioning

Prof. Bachelor	PB1	Office Management-languages (accountancy, taxation, library, marketing, communication policy)
	PB2	Industrial sciences (audiovisual technique, photography, graphical designer, (bio)chemistry, electricity, electromechanics, telecommunication, image forming, textiles)
	PB3	Computer Science (A1)
	PB4	welfare assistant, psychologic assistant, medical pedagogy
	PB5	Teacher in lower secondary education
	PB6	Nursing, midwife
	PB7	(Kindergarden) teacher
	PB8	Health care (laboratory assistant, speech therapy, physical therapy, ergotherapy, dietetics)
	PB9	Architecture
	PB10	Others
	PB11	Tourism
	PB12	Catering industry
	PB13	Legal
	PB14	Construction
	PB15	Agriculture, horticulture
	PB16	Finances, Insurances, Real estate
	PB17	Media, Advertising
	PB18	Logistics, Transportation
College Master	CM1	Industrial Engineer
	CM2	Public administration, commercial engineer, commercial sciences
	CM3	Journalism, interpreter, translator
	CM4	Master in arts (drama, music, design, expressive arts)
	CM5	Others
	CM6	Architect
University	U1	Applied economic sciences
	U2	Civil engineer, civil engineer architect
	U3	Social sciences (sociology, political sciences, communication sciences, administrative sciences)
	U4	Physical sciences (biology, mathematics, chemics, physics, geology, geography)
	U5	Commercial engineer, commercial engineer policy informatics
	U6	Law
	U7	Languages, philology
	U8	Psychology, pedagogy
	U9	Agricultural engineer
	U10	(Art) history, music science, archeology
	U11	Computer Science
	U12	Medicine (doctor, dentist)
	U13	Criminology
	U14	Physical education, physical therapy, speech therapy
	U15	Pharmaceutical sciences
	U16	Economic sciences
	U17	Others
	U18	Philosophy, theology
	U19	Veterinary science
	U20	Medical-social sciences, hospital sciences
	U21	Bio-medical sciences
	U22	Military sciences

Table 4: Sectoral Partitioning

Primary sector	PR1	Primary (Agriculture, stock farming, horticulture, fishing, farming)
Industry	I1	Processing of metallic products, construction of machinery, electrotechnic industry, car assembly
	I2	Chemical industry, oil and gas, processing of rubber and plastics
	I3	Pharmaceutical industry
	I4	Production of food, drinks and smoke products
	I5	Construction, road construction
	I6	Timber industry, paper industry, graphic industry, production of glass, bricks and cement
	I7	Textiles, shoe and clothing industry, leather industry
	I8	Production and distribution of energy, water supply
Services	S1	IT
	S2	Banking and insurance
	S3	Consulting, business services, research & development
	S4	Transport, logistics en distribution
	S5	Retail, wholesale
	S6	Telecommunication
	S7	Advertising and media, entertainment and communication
	S8	Services regarding human resources (eg. Selection, education, temporary employment agency)
	S9	Tourism & leisure
	S10	Catering
	S11	Other services to firms (security, cleaning, leasing, maintenance)
Public sector	P1	Health service
	P2	Education
	P3	Federal government
	P4	Welfare, community services
	P5	Local governments (community, province)
	P6	Governments of districts and communities
	P7	Socio-cultural sector
	P8	International governments and institutions
Others	Z1	Others